Bernd SCHUERMANN et al.
Docket No. 1454.1685

MARKED-UP COPY OF SUBSTITUTE SPECIFICATION

~~Description~~ TITLE OF THE INVENTION

IDENTIFICATION OF PHARMACEUTICAL TARGETS

CROSS REFERENCE TO RELATED APPLICATIONS

[0001] This application is based on and hereby claims priority to PCT Application No. PCT/EP2004/051835 filed on August 18, 2004 and German Application 10342274.9 filed September 12, 2003, the contents of which are hereby incorporated by reference.

BACKGROUND OF THE INVENTION

[0002] It is estimated that the human genome comprises 20,000 to 80,000 genes, which contain the genetic code for around one million proteins. In the specialized body cells only subsets of all genes are actually read off (expressed) in each case. The totality of the proteins created in this way is referred to as the proteome of this cell. The interplay of the proteins with each other as well as with the DNA represents the most important part of the machinery which underlies the development of the human body from the fertilized egg cell as well as all body functions. From the information technology standpoint the genome thereby represents a procedural code for the structure and function of the human body.

[0003] Many illnesses and malfunctions of the body are attributable to faults in the functional network of genome and proteome. Thus ~~a number of~~a plurality of medicaments operate as agonists or antagonists of specific target proteins, i.e. they strengthen or weaken the function of a protein with the aim of returning the function of the regulatory network formed from proteome and genome back into a normal functional mode. These targets have previously been derived using heuristic principles from biochemical considerations. It is often unclear in such cases whether the malfunction of a protein represents the actual cause of the illness or only one of the symptoms of a hidden mis-regulation at another point of the network.

[0004] The simulation of nerve cells (neurons) and their biological functionality by artificial neurons is known from Zell, A., "Simulation Neuronaler networks" (Simulation of Neuronal Networks), P. 35 to 51, P. 55 to 86, Addison-Wesley Longman Verlag GmbH, 1994, 3rd

unamended reprint, R. Oldenbourg Verlag, ISBN 3-486-24350-0, 2000 ("Zell pages 35 to 51 and 55-86")[~~1~~] and Patent Application PCT/DE02/03381 ("PCT03381")[~~5~~].

[0005] Making a distinction between two types of synapses or nerve cells, an excitatory synapse 251 or nerve cell and an inhibiting synapse 252 or nerve cell is further known from Zell pages 35 to 51 and 55-86[~~1~~].

[0006] Inhibitory synapses 252 reduce electrical potentials to be transmitted or forwarded, excitatory synapses 251 increase electrical potentials to be transmitted.

[0007] Further information about the structure and functionality of a nerve cell as well as for nerve conductance is given in Zell pages 35 to 51 and 55-86[~~1~~].

[0008] Furthermore an artificial nerve cell (artificial neuron) which emulates a (biological) nerve cell is known from Zell pages 35 to 51 and 55-86[~~1~~] and [~~5~~]PCT03381.

[0009] At first glance such an artificial neuron is a mathematical mapping, which, in accordance with the transmission behavior of the biological nerve cell, maps an input variable of the artificial neuron onto an output variable of the artificial neuron.

[0010] In compliance with the biological template an artificial neuron ~~consists of~~ has three components: the cell body, the dendrites, which sum input signals into the artificial neuron, the axon, which forwards the output signal of the artificial neuron to the outside, branches and comes into contact with the dendrites of subsequent artificial neurons via synapses.

[0011] A strength of a synapse or the type of the synapse is mostly represented by a numeric value or by its leading sign. This value is referred to as the connection weight.

[0012] In accordance with the biological template a transmission behavior or the mapping behavior of an artificial neuron can be mapped as described in Zell pages 35 to 51 and 55-86[~~1~~] and [~~5~~]PCT03381.

[0013] Further information about artificial neurons and their functionality is given in Zell pages 35 to 51 and 55-86[~~1~~] and [~~5~~]PCT03381.

[0014] The linkage of individual neurons with each other is further known from Zell, A., "Simulation Neuronaler networks" (Simulation of Neuronal Networks), P. 87 to 114, Addison-

Wesley Longman Verlag GmbH, 1994, 3rd unamended reprint, R. Oldenbourg Verlag, ISBN 3-48624350-0, 2000 ("Zell pages 87-114")[2] and [5]PCT03381. Such an arrangement with linked neurons is referred to as a neuronal network. The basics of neuronal networks, for example different types of neuronal networks, training methods for neuronal networks, references to biological nerve cell arrangements, are described in [2]Zell pages 87-114.

[0015] An application of a mean-field model to the description of a complex system is known from J. J. Binney, N. J. Dowrick, A. J. Fisher, M. E. J. Newman, "The Theory of Critical Phenomena", Chap. 6: Mean-Field Theory, Clarendon Press Oxford 1992[3] and [5]PCT03381. With a mean-field model stochastic interaction influences between components of a system are approximated by a mean interaction influence. This allows stochastic systems which cannot be described analytically to be reduced to describable, deterministic systems.

[0016] The application of the mean-field model to the description of a neuron structure is known from C. Koch, I. Segev (Hrsg), "Methods of Neural Modeling: From Ions To Networks", Chap 13: D. Hansel and H. Sompolinsky: "Modeling Feature Selectivity in Local Circuits", MIT Press, Cambridge, 1998[4] and [5]PCT03381.

SUMMARY OF THE INVENTION

[0017] One possible ~~The~~ object of the invention is to improve the identification of proteins which are suitable as targets for treatment without medicaments of genetic diseases or problems.

[0018] ~~This object is achieved by the inventions in accordance with the independent claims. Advantageous developments of the inventions are identified in the subclaims. In accordance with the invention~~ The inventors propose to determine a plurality of gene expression patterns of similar types of cells or of a tissue ~~is determined~~, with an expression rate of the genes of the cell being determined in each case. The plurality of gene expression patterns is determined such that the chronological sequence of the gene expression pattern of the cell can be at least partly reconstructed. A dynamic model of the regulatory network made up of genome and proteome of the cell is formed by forming an equivalent neuronal network in the following way:

i) A gene of the genome as well as the associated protein are represented by a neuron of the equivalent neuronal network.

ii) The expression rate of a gene is represented by a non-negative activity of the equivalent neuron.

iii) The regulatory effect of the protein on a gene is represented by a synaptic connection from the neuron equivalent to the protein to the neuron equivalent to the gene.

iv) The type of regulatory effect (strengthening or inhibiting) is represented in the neuronal network by the leading sign and the strength of the associated synaptic weight.

v) In a further development it is also possible to represent a post-translational modification of a first protein by a second protein by a synaptic connection with multiplicative effect from the second neuron to the first neuron.

vi) In a further development an external influence on the regulatory network can be represented by an input node of the equivalent neuronal network.

[0019] The equivalent neuronal network is compared with the gene expression patterns determined and is adapted to these. From the adapted neuronal network the regulatory network of the cell investigated is deduced.

[0020] There is thus an equivalence relationship between the functional network of the genome and proteome on the one side and the neuronal network of the human brain on the other side, which both represent strongly networked closed-loop systems. This mapping brings about successful modeling of the functional network of proteins and genes.

[0021] The method allows the identification of target proteins on a systematic basis. The equivalence relationship described can be established between genetic and neuronal networks. The dynamic interaction of genes and regulatory proteins is thus modeled by a dynamic neuronal network. The method uses the information contained in the chronological sequence of the gene expression pattern for the identification of causal regulatory interrelationships.

[0022] As a rule genetic illnesses lead to complex malfunctions which however often only lead back to a few malfunctioning genes or proteins. Until now these key genes have not been known except in individual cases. Instead heuristic processes have been used in conventional target finding to search for targets for which regulation without medicament would restore the healthy gene expression pattern in the best possible way.

[0023] Recent estimates talk of 10,000 different proteins as possible targets in the human genome which it would not be practicable to thin out using a heuristic approach alone.

[0024] The model approach described here represents a powerful method for systematic target finding, that its for identification of one or more key genes or proteins which are located at the start of the regulation cascade and which for example introduce an organic development, regulate the regeneration capability of tissue but which are also responsible for mostly complex changes of gene expression patterns in the event of illness.

[0025] The method described allows a computer-based target finding which is able to analyze the large amount of data and the numerous and complex interrelationships.

[0026] It allows the following application areas to be specified:

[0027] - Model-based support of research activities for decoding the human morphogenesis and there by the general principles of genetically controlled growth, regeneration and breakdown processes.

[0028] - Support for the identification of target proteins which are fundamentally responsible for problems with growths e.g. unrestricted tumor growth and do not just represent one symptom. Novel methods for highly-sensitive early tumor diagnosis can be derived from this but also treatment methods such as selectively induced cell death (apoptosis) in tumor cells.

[0029] - Support for the identification of regulatory proteins which intervene into growth and regeneration processes. This would overcome a significant hurdle on the way to induced regeneration of organs.

BRIEF DESCRIPTION OF THE DRAWINGS

[0030] These and other objects and advantages of the present invention will become more apparent and more readily appreciated from the following description of the preferred embodiments, taken in conjunction with the accompanying drawings of which:~~The invention is explained in more detail below using exemplary embodiments which are shown schematically in the Figures. The same reference numbers in individual Figures identify the same elements in each case. In detail the Figures show:~~

Fig. 1 a schematic diagram of the regulatory processes which determine the

expression pattern of a cell;

Fig. 2 a schematic diagram of the networking of neurons;

Fig. 3 the potentials within a dendrite or a neuron as a function of the time and

Fig. 4 a schematic diagram or a modulatory synapse.

DETAILED DESCRIPTION OF THE PREFERRED EMBODIMENT

[0031] Reference will now be made in detail to the preferred embodiments of the present invention, examples of which are illustrated in the accompanying drawings, wherein like reference numerals refer to like elements throughout.

[0032] Fig. 1 illustrates the main interactions between genes and proteins of a section of DNA. The interactions are included as the basis for the description of the genomic regulatory network.

[0033] The top part of Fig 1 shows schematically an external signal affecting the cell from the outside - within the framework of intercellular communication for example - which is accepted for example by a transmembrane receptor protein (e.g. from a calcium channel) and is transmitted in an appropriate manner into the inside of the cell, initiates the production of the genes A, B, C and D of the DNA section.

[0034] The option thus basically exists for influencing the expression rate of individual genes of a cell over the path mentioned from outside the cells.

[0035] A not necessarily contiguous section of the DNA is referred to as a gene which contains the genetic code for a protein or also for a group of proteins. In general the DNA features what are known as exons and introns. Exons represent parts of the DNA which actually encode a protein. Introns represent parts of the DNA which do not directly encode a protein. In a first approximation they have no function. Exons and introns alternate with each other in the DNA. If a gene is referred to as the quantity of the exons which together encode a specific protein, such a gene - as mentioned above - is as a rule not contiguous.

[0036] The production process of a protein from a gene, for example protein A, starting from gene A in Fig. 1, is referred to as the expression of this gene. The conversion of the DNA code

of the gene into the chain of the amino acids of the protein is referred to as translation. The rate at which the protein A is produced in a given context is called its expression rate.

[0037] Not all genes are expressed in a cell. Instead different cell types are differentiated by their gene expression pattern. This then often also applies to the difference between diseased and healthy cells.

[0038] The expression pattern of a cell is determined by the regulatory processes shown schematically in Fig. 1. The regulatory processes are essentially determined by a few important interactions between proteins and genes as well as between the proteins themselves.

[0039] Thus the expression rate of a gene A can be regulated by the presence of another protein B, i.e. increased, reduced or brought to a standstill. In this example the protein B acts in a regulatory way on the gene A or the protein A. The protein components of activator complexes can for example be reckoned to be regulatory proteins. Regulatory proteins can operate on many target genes simultaneously.

[0040] A second type of interaction ~~consists of~~ is a post-translational modification of proteins, i.e. the modification of proteins after their translation. As a rule the post-translational modification of a protein occurs directly after the translation, i.e. before the protein acts in the cell. Thus for example many proteins are phosphorylized or glycolyzed by specific enzymes, i.e. the target protein is put into its functional state by appending or splitting off chemical groups or is put into a state of in which it no longer has any effect. Post-translational modification can thus temporarily switch on or switch off the functions of a protein where necessary.

[0041] In Fig. 1 the protein A is what is referred to as an effector protein, i.e. it operates within the cell on other substances and not directly on the genome or the proteome. In Fig. 1 the protein C during the course of post-translational modification modifies the function of the effector protein A.

[0042] Protein B is a regulatory protein since it determines the expression rate of the protein A by interacting with that DNA section which contains the gene A. The protein D thus modifies the function of a regulatory protein (protein B) in the course of the post-translational modification.

[0043] Database

[0044] The nucleic acid sequence of the human DNA is very largely known. The genes encoded by the DNA have also been identified to an increasing extent. Not quite so complete is the knowledge about the proteins including the post translational modified proteins possibly produced by interactions between the proteins. In any event new sequencing and high-throughput screening processes allow further genes and proteins to be quickly identified.

[0045] A further important step for clarifying the expression pattern of a cell has been completed with the development of high-throughput hybridization techniques. With this method the expression rate of many 100 different genes is tested simultaneously on what is known as a microarray. With the aid of this method it is possible to determine the gene expression pattern of a cell.

[0046] To this end the mRNA (messenger RNAs) synthesized in the cell are determined as a rule. The mRNA is an intermediate product in the translation of the gene into a protein. The mRNA is thus a preliminary stage in the formation of the protein and points to the formation of the associated protein. The cell to be investigated is first isolated. Subsequently it is deduced. Suitable rationalization steps are used to isolate the mRNAs from the cell. Then the mRNA is translated using reverse transcriptasis into cDNA (complementary DNA). This is generally amplified by linear PCR (polymerase chain reaction). The cDNA thus obtained is analyzed with the aid of suitable microarrays, e.g. DNA chips, qualitatively or quantitatively. With modern microarrays the expression rates of 5,000 and more genes can be calibrated simultaneously.

[0047] Because of these improved techniques there is now comprehensive knowledge about the human genome and proteome as well as about the interactions between proteins and genes or between the proteins themselves.

[0048] Of particular importance are data records in which the chronological sequence of gene expression patterns in a tissue is stored. What are known as longitudinal hybridization studies, i.e. chronological sequences of gene expression patterns during the organ differentiation as part of the embryonal development is one example that might be mentioned. Time-resolved gene expression data also exists for the cell division cycle of single cell creatures and is also possible for more complex tissue.

[0049] Neuronal modeling of genome and proteome (cf. ~~[5]~~PCT03381)

[0050] A general outline of the modeling principle is given below. The basics are known from ~~[5]~~PCT03381.

[0051] The basic principle ~~consists of~~ relates to establishing an equivalence relationship between the functional network of the genome and proteome on the one hand and the neuronal network of the human brain on the other hand, which both represent heavily networked closed-loop systems.

[0052] The neuronal network of the human brain is illustrated below in ~~a number of~~a plurality of fundamentals with reference to Fig. 2.

[0053] In the human brain there are around 100 billion nerve cells 20 (neurons), which each exchange information with tens of thousands of other nerve cells 20. The information passes from a neuron 20 via the axon 22 belonging to each neuron to another neuron 20. Each neuron has precisely one axon to send information to other neurons. In its further progress and axon typically branches around one thousand times, so that a neuron 20 can send information of via its axon 22 to around one thousand other neurons 20.

[0054] To receive information neurons 20 have dendrites 24. The axon 22 carrying information is connected via a synapse 26 with the dendrites 24. The information passes via this synapse 26 from the axon 22 into the dendrite 24 and thereby from the emitting neuron 22 to the downstream neuron. Between thousands and hundreds of thousands of axons 22 or synapses 26 can access a single dendrite 24 so that a downstream neuron 20 can receive signals from many 1000 upstream neurons.

[0055] Reference is made to Fig. 3 below which shows the potentials within a dendrite or a neuron as a function of time. The information is exchanged between the neurons 20 in the form of action potentials (spikes) 30, which each neuron 20 emits via its axon 22. The spikes evoke renewed signals in the downstream neurons 20, the so-called post-synaptic potentials (PSPs) 32. The size of these PSPs depends on the transmission strength or the synaptic weight w of the synapse concerned. .

[0056] The text below refers to Fig.4. Fig. 4 shows dendrite 24, to which a first synapse 26 couples. A second synapse 36 is coupled to this first synapse 26. This second synapse 36 is called a modulatory synapse. If, with reference to Fig. 3 we designate the post-synaptic

potential 32 as PSP which would form in dendrite 24 as a result of the effect of the first synapse 26 in the absence of the modulatory synapse 36, this can be represented by

$$PSP=w\cdot\varepsilon(t),$$

[0057] with w, as defined above, representing the synaptic weight of the first synapse 26 and 8(t) the timing of the post-synaptic potential 32 in a suitable normalization.

[0058] If in addition the modulatory synapse 36 accesses the first synapse 26, this produces a modified post-synaptic potential PSP' in the dendrite 24 which can be expressed by a multiplicative term act:

$$PSP'=act\cdot w\cdot\varepsilon(t)=act\cdot PSP.$$

[0059] In this case act identifies the activity of the modulatory synapse 36.

[0060] For example dopaminergic synapses have a modulatory character in the central nervous system, that is in the brain and the spinal cord.

[0061] The neuronal activity of each neuron, i.e. the number of the spikes emitted for each unit of time, is produced - in simple terms - by a non-linear and chronologically non-local function of all incoming post-synaptic potentials. If this function exceeds a specific threshold value a spike 30 is initiated and transmitted via the axon 22.

[0062] Thus the biological neuronal network of the brain represents a complex non-linear system which also features a high networking density. To describe this system in a formal model neuro-information technology has developed powerful theories and algorithms in recent years (e.g. compartment model, spike response model, mean-field model, multi-modular neuro cognitive model, Bayes belief networks).

[0063] These theories or equations correspond in their structure to the equations derived above for reaction kinetics. Thus the regulatory network of genome and proteome of a cell can be mapped to an equivalent neuronal network as follows:

[0064] -A gene A of the genome (understood here as that combination of exons which uniquely encode a protein) as well as the associated protein A are identified with a neuron A of the equivalent neuronal network. Since in the gene expression pattern only the mRNAs or

cDNAs are qualitatively analyzed, it is also not possible at the level of the gene expression pattern to distinguish between genes and proteins just like that.

[0065] -The expression rate of a gene A is expressed as a non-negative activity, e.g. the spike rate of the neuron A.

[0066] -If a protein B acts in a regulatory manner on a downstream gene A, the equivalent neuronal network contains a synaptic connection from neuron B to the equivalent downstream neuron A.

[0067] - The type of regulatory effect (strengthening or inhibiting) is specified in the neuronal network by the leading sign and the strength of the associated synaptic weight.

[0068] - A post-translational modification of a protein by another protein, in Fig. 1 for example the modification of protein B by protein D, corresponds to the effect of a modulatory synapse in the central nervous system. Modulatory synapses are described in artificial neuronal networks by synaptic connections with multiplicative effect on other synapses. The equivalent reflection of a post-translational modification of the protein B by a protein D is thus a synaptic connection with multiplicative effect from neuron D to neuron B.

[0069] - External signals are identified by input nodes of the equivalent neuronal network.

[0070] The equivalence relationship described can be established between genetic and neuronal networks. The dynamic interaction of genes and regulatory proteins is thus modeled by a dynamic neuronal network.

[0071] Networks of spiking neurons count as suitable neuronal algorithms but also mean-field models which take into account the explicit passage of time of the signal transmission between the neurons by the explicit description of the post-synaptic potentials. They allow the modeling of the development over time of the neuronal activities in the network as a result of external stimulation or intrinsic activity.

[0072] The development over time of the concentrations which is produced by the reaction kinetics between the molecules involved (e.g. between regulatory protein and DNA promoter) will thus be replaced by the time sequence of the activities of the neurons so that the resulting

network model for simulating the timing development of gene expression patterns can be included.

[0073] The neuronal activities over time can be included for this type of neuronal network. Since the neuronal activity corresponds to the gene expression patterns, the two can be compared to each other. The neuronal network corresponds to a simulated gene expression pattern.

[0074] The object of the modeling is to determine the regulatory network underlying the expression sequence, i.e. to answer the following question: "Which neuronal networking structure with which weights and reaction constants is consistent with the observed gene expression sequence?"

[0075] To answer this question the network is trained with a method oriented to structured learning: An attempt is made to explain the observed behavior with as few regulatory connections as possible but also as well as possible, that is to find the simplest model consistent with the data.

[0076] A preferred optimization method minimizes the total deviation between measured and simulated gene expression patterns by using a "sparse prior", that is an additional condition which penalizes the co-existence of many connections with small weights in favor of fewer regulatory connections. An option for implementing such a sparse prior is known to those skilled in the art.

[0077] Cross-validation and statistical optimization allow the uniqueness of the solution to be estimated as well as its ability to predict (generalization capability).

[0078] Causal relationships between genes but also the role of different genes can be taken from the trained network on the basis of the connection structure of the neurons. Thus an asymmetrical weight only from gene B to gene A indicates that gene B regulates gene A. At the same time in the model different genes or regulatory connections can be artificially switched off or switched on and the effects of the gene expression pattern with the target quantified, to identify the cause(s) of the illness-related changes of the gene expression (known as inverse modeling).

[0079] The invention has been described in detail with particular reference to preferred embodiments thereof and examples, but it will be understood that variations and modifications can be effected within the spirit and scope of the invention covered by the claims which may include the phrase "at least one of A, B and C" as an alternative expression that means one or more of A, B and C may be used, contrary to the holding in *Superguide v. DIRECTV*, 69 USPQ2d 1865 (Fed. Cir. 2004).